

TEXT MINING SYSTEM FOR WEB-BASED BUSINESS INTELLIGENCE

RELATED PATENT APPLICATIONS

1a1

This application ~~claims the benefit of U.S.~~
Provisional Application No. 60/206,772, filed May 25,
2000 and entitled "Web-Based Customer Lead Generator".

- 5 The present patent application and additionally the
following patent applications are ~~each conversions~~ from
the foregoing provisional filing: Patent Application
Serial No. _____ (Attorney Docket No.
068082.0105) entitled "Web-Based Customer Lead Generator
10 System" and filed May 21, 2001; Patent Application Serial
No. _____ (Attorney Docket No. 068082.0114)
entitled "Web-Based Customer Prospects Harvester System"
and filed May 21, 2001; Patent Application Serial No.
_____ (Attorney Docket No. 068082.0111) entitled
15 "Database Server System for Web-Based Business
Intelligence" and filed _____; Patent
Application Serial No. _____ (Attorney Docket No.
068082.0112) entitled "Data Mining System for Web-Based
Business Intelligence" and filed _____; Patent
20 Application Serial No. _____ (Attorney Docket No.
068082.0115) entitled "Text Indexing System for Web-Based
Business Intelligence" and filed _____.

TECHNICAL FIELD OF THE INVENTION

This invention relates to electronic commerce, and more particularly to business intelligence software tools for acquiring leads for prospective customers, using

5 Internet data sources.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2

BACKGROUND OF THE INVENTION

Most small and medium sized companies face similar challenges in developing successful marketing and sales campaigns. These challenges include locating qualified prospects who are making immediate buying decisions. It is desirable to personalize marketing and sales information to match those prospects, and to deliver the marketing and sales information in a timely and compelling manner. Other challenges are to assess current customers to determine which customer profile produces the highest net revenue, then to use those profiles to maximize prospecting results. Further challenges are to monitor the sales cycle for opportunities and inefficiencies, and to relate those findings to net revenue numbers.

Today's corporations are experiencing exponential growth to the extent that the volume and variety of business information collected and accumulated is overwhelming. Further, this information is found in disparate locations and formats. Finally, even if the individual data bases and information sources are successfully tapped, the output and reports may be little more than spreadsheets, pie charts and bar charts that do not directly relate the exposed business intelligence to the companies' processes, expenses, and to its net revenues.

With the growth of the Internet, one trend in developing marketing and sales campaigns is to gather customer information by accessing Internet data sources. Internet data intelligence and data mining products face specific challenges. First, they tend to be designed for

use by technicians, and are not flexible or intuitive in their operation; secondly, the technologies behind the various engines are changing rapidly to take advantage of advances in hardware and software, and finally, the

5 results of their harvesting and mining are not typically related to a specific department goals and objectives.

SUMMARY OF THE INVENTION

One aspect of the invention is a text mining system for collecting business intelligence about a client, as well as for identifying prospective customers of the client. The text mining system is used in a lead generation system accessible by the client via the Internet.

The mining system has various components, including a data acquisition process that extracts textual data from various Internet sources, a database for storing the extracted data, a text mining server that executes query-based searches of the database, and an output repository. A web server provides client access to the repository, and to the mining server.

15

BRIEF DESCRIPTION OF THE DRAWINGS

FIGURE 1 illustrates the operating environment for a web based lead generator system in accordance with the invention.

5 FIGURE 2 illustrates the various functional elements of the lead generator system.

FIGURE 3 illustrates the various data sources and a first embodiment of the prospects harvester.

10 FIGURES 4 and 5 illustrate a database server system, which may be used within the lead generation system of FIGURES 1 and 2.

FIGURES 6 and 7 illustrate a data mining system, which may be used within the lead generation system of FIGURES 1 and 2.

15 FIGURES 8 and 9 illustrate a text mining system, which may be used within the lead generation system of FIGURES 1 and 2.

20 FIGURES 10 and 11 illustrate a text indexing system, which may be used within the lead generation system of FIGURES 1 and 2.

FIGURE 12 illustrates a digital voice recording mining system, which may be used within the lead generation system of FIGURES 1 and 2.

DETAILED DESCRIPTION OF THE INVENTION

Lead Generator System Overview

FIGURE 1 illustrates the operating environment for a web-based customer lead generation system 10 in accordance with the invention. System 10 is in communication, via the Internet, with unstructured data sources 11, an administrator 12, client systems 13, reverse look-up sources 14, and client applications 15.

The users of system 10 may be any business entity that desires to conduct more effective marketing campaigns. These users may be direct marketers who wish to maximizing the effectiveness of direct sales calls, or e-commerce web site who wish to build audiences.

In general, system 10 may be described as a web-based Application Service Provider (ASP) data collection tool. The general purpose of system 10 is to analyze a client's marketing and sales cycle in order to reveal inefficiencies and opportunities, then to relate those discoveries to net revenue estimates. Part of the latter process is proactively harvesting prequalified leads from external and internal data sources. As explained below, system 10 implements an automated process of vertical industry intelligence building that involves automated reverse lookup of contact information using an email address and key phrase highlighting based on business rules and search criteria.

More specifically, system 10 performs the following tasks:

- Uses client-provided criteria to search Internet postings for prospects who are discussing products or services that are related to the client's business offerings
- 5 • Selects those prospects matching the client's criteria
- Pushes the harvested prospect contact information to the client, with a link to the original document that verifies the prospects interest
- Automatically opens or generates personalized sales
10 scripts and direct marketing materials that appeal to the prospects' stated or implied interests
- Examines internal sales and marketing materials, and by applying data and text mining analytical tools, generates profiles of the client's most profitable
15 customers
- Cross-references and matches the customer profiles with harvested leads to facilitate more efficient harvesting and sales presentations
- In the audience building environment, requests
20 permission to contact the prospect to offer discounts on services or products that are directly or indirectly related to the conversation topic, or to direct the prospect to a commerce source.

System 10 provides open access to its web site. A
25 firewall (not shown) is used to prevent access to client records and the entire database server. Further details of system security are discussed below in connection with FIGURE 5.

Consistent with the ASP architecture of system 10,
30 interactions between client system 13 and system 10 will typically be by means of Internet access, such as by a

web portal. Authorized client personnel will be able to create and modify profiles that will be used to search designated web sites and other selected sources for relevant prospects.

- 5 Client system 11 may be any computer station or network of computers having data communication to lead generator system 10. Each client system 11 is programmed such that each client has the following capabilities: a master user account and multiple sub user accounts, a
- 10 user activity log in the system database, the ability to customize and personalize the workspace; configurable, tiered user access; online signup, configuration and modification, sales territory configuration and representation, goals and target establishment, and
- 15 online reporting comparing goals to target (e.g., expense/revenue; budget/actual).

- Administration system 14 performs such tasks as account activation, security administration, performance monitoring and reporting, assignment of master userid and
- 20 licensing limits (user seats, access, etc.), billing limits and profile, account termination and lockout, and a help system and client communication.

- System 10 interfaces with various client applications 15. For example, system 10 may interface
- 25 with commercially available enterprise resource planning (ERP), sales force automation (SFA), call center, e-commerce, data warehousing, and custom and legacy applications.

Lead Generator System Architecture

FIGURE 2 illustrates the various functional elements of lead generator system 10. In the embodiment of FIGURE 2, the above described functions of system 10 are
5 partitioned between two distinct processes.

A prospects harvester process 21 uses a combination of external data sources, client internal data sources and user-parameter extraction interfaces, in conjunction with a search, recognition and retrieval system, to
10 harvest contact information from the web and return it to a staging data base 22. In general, process 21 collects business intelligence data from both inside the client's organization and outside the organization. The information collected can be either structured data as in
15 corporate databases/spreadsheet files or unstructured data as in textual files.

Process 21 may be further programmed to validate and enhance the data, utilizing a system of lookup, reverse lookup and comparative methodologies that maximize the
20 value of the contact information. Process 21 may be used to elicit the prospect's permission to be contacted. The prospect's name and email address are linked to and delivered with ancillary information to facilitate both a more efficient sales call and a tailored e-commerce sales
25 process. The related information may include the prospect's email address, Web site address and other contact information. In addition, prospects are linked to timely documents on the Internet that verify and highlight the reason(s) that they are in fact a viable
30 prospect. For example, process 21 may link the contact data, via the Internet, to a related document wherein the

contact's comments and questions verify the high level value of the contact to the user of this system (the client).

A profiles generation process 25 analyzes the user's in-house files and records related to the user's existing customers to identify and group those customers into profile categories based on the customer's buying patterns and purchasing volumes. The patterns and purchasing volumes of the existing customers are overlaid on the salient contact information previously harvested to allow the aggregation of the revenue-based leads into prioritized demand generation sets. Process 25 uses an analysis engine and both data and text mining engines to mine a company's internal client records, digital voice records, accounting records, contact management information and other internal files. It creates a profile of the most profitable customers, reveals additional prospecting opportunities, and enables sales cycle improvements. Profiles include items such as purchasing criteria, buying cycles and trends, cross-selling and up-selling opportunities, and effort to expense/revenue correlations. The resulting profiles are then overlaid on the data obtained by process 21 to facilitate more accurate revenue projections and to enhance the sales and marketing process. The client may add certain value judgments (rankings) in a table that is linked to a unique lead id that can subsequently be analyzed by data mining or OLAP analytical tools. The results are stored in the deliverable database 24.

Profiles generation process 25 can be used to create a user (client) profiles database 26, which stores

profiles of the client and its customers. As explained below, this database 26 may be accessed during various data and text mining processes to better identify prospective customers of the client.

5 Web server 29 provides the interface between the client systems 13 and the lead generation system 10. As explained below, it may route different types of requests to different sub processes within system 10. The various web servers described below in connection with FIGURES 4-
10 11 may be implemented as separate servers in communication with a front end server 29. Alternatively, the server functions could be integrated or partitioned in other ways.

Data Sources

15 FIGURE 3 provides additional detail of the data sources of FIGURES 1 and 2. Access to data sources may be provided by various text mining tools, such as by the crawler process 31 or 41 of FIGURES 3 and 4.

One data source is newsgroups, such as USENET. To
20 access discussion documents from USENET newsgroups such as "news.giganews.com", NNTP protocol is used by the crawler process to talk to USENET news server such as "news.giganews.com." Most of the news servers only archive news articles for a limited period (giganews.com
25 archives news articles for two weeks), it is necessary for the iNet Crawler to incrementally download and archive these newsgroups periodically in a scheduled sequence. This aspect of crawler process 31 is controlled by user-specified parameters such as news
30 server name, IP address, newsgroup name and download frequency, etc.

Another data source is web-Based discussion forums. The crawler process follows the hyper links on a web-based discussion forum, traverse these links to user or design specified depths and subsequently access and
5 retrieve discussion documents. Unless the discussion documents are archived historically on the web site, the crawler process will download and archive a copy for each of the individual documents in a file repository. If the discussion forum is membership-based, the crawler process
10 will act on behalf of the authorized user to logon to the site automatically in order to retrieve documents. This function of the crawler process is controlled by user specified parameters such as a discussion forum's URL, starting page, the number of traversal levels and
15 crawling frequency.

A third data source is Internet-based or facilitated mailing lists wherein individuals send to a centralized location emails that are then viewed and/or responded to by members of a particular group. Once a suitable list
20 has been identified a subscription request is initiated. Once approved, these emails are sent to a mail server where they are downloaded, stored in system 10 and then processed in a fashion similar to documents harvested from other sources. The system stores in a database the
25 filters, original URL and approval information to ensure only authorized messages are actually processed by system 10.

A fourth data source is corporations' internal documents. These internal documents may include sales
30 notes, customer support notes and knowledge base. The crawler process accesses corporations' internal documents

from their Intranet through Unix/Windows file system or alternately be able to access their internal documents by riding in the databases through an ODBC connection. If internal documents are password-protected, crawler
5 process 31 acts on behalf of the authorized user to logon to the file systems or databases and be able to subsequently retrieve documents. This function of the crawler process is controlled by user-specified parameters such as directory path and database ODBC path,
10 starting file id and ending file id, and access frequency. Other internal sources are customer information, sales records, accounting records, and digitally recorded correspondence such as e-mail files or digital voice records.

15 A fifth data source is web pages from Internet web sites. This function of the crawler process is similar to the functionality associated with web-discussion-forums. Searches are controlled by user-specified parameters such as web site URL, starting page, the
20 number of traversal levels and crawling frequency.

Database Server System

FIGURES 4 and 5 illustrate a database server system 41, which may be used within system 10 of FIGURES 1 and 2. FIGURE 4 illustrates the elements of system 41 and
25 FIGURE 5 is a data flow diagram. Specifically, system 41 could be used to implement the profiles generation process 25, which collects profile data about the client.

The input data 42 can be the client's sales data, customer-contact data, customer purchase data and account
30 data etc. Various data sources for customer data can be contact management software packages such as ACT,

MarketForce, Goldmine, and Remedy. Various data sources for accounting data are Great Plains, Solomon and other accounting packages typically found in small and medium-sized businesses. If the client has ERP (enterprise
5 resource planning) systems (such as JD Edwards, PeopleSoft and SAP) installed, the data sources for customer and accounting data will be extracted from ERP customer and accounting modules. This data is typically structured and stored in flat files or relational
10 databases. System 41 is typically an OLAP (On-line analytic processing) type server-based system. It has five major components. A data acquisition component 41a collects and extracts data from different data sources, applying appropriate transformation, aggregation and
15 cleansing to the data collected. This component consists of predefined data conversions to accomplish most commonly used data transformations, for as many different types of data sources as possible. For data sources not covered by these predefined conversions, custom
20 conversions need to be developed. The tools for data acquisition may be commercially available tools, such as Data Junction, ETI*EXTRACT, or equivalents. Open standards and APIs will permit employing the tool that affords the most efficient data acquisition and migration
25 based on the organizational architecture.

Data mart 41b captures and stores an enterprise's sales information. The sales data collected from data acquisition component 41a are "sliced and diced" into
30 multidimensional tables by time dimension, region dimension, product dimension and customer dimension, etc. The general design of the data mart follows data

warehouse/data mart Star-Schema methodology. The total number of dimension tables and fact tables will vary from customer to customer, but data mart 41b is designed to accommodate the data collected from the majority of
5 commonly used software packages such as PeopleSoft or Great Plains.

Various commercially available software packages, such as Cognos, Brio, Informatica, may be used to design and deploy data mart 41b. The Data Mart can reside in
10 DB2, Oracle, Sybase, MS SQL server, P.SQL or similar database application. Data mart 41b stores sales and accounting fact and dimension tables that will accommodate the data extracted from the majority of industry accounting and customer contact software
15 packages.

A Predefined Query Repository Component 41c is the central storage for predefined queries. These predefined queries are parameterized macros/business rules that extract information from fact tables or dimension tables
20 in the data mart 41b. The results of these queries are delivered as business charts (such as bar charts or pie charts) in a web browser environment to the end users. Charts in the same category are bounded with the same predefined query using different parameters. (i.e.
25 quarterly revenue charts are all associated with the same predefined quarterly revenue query, the parameters passed are the specific region, the specific year and the specific quarter). These queries are stored in either flat file format or as a text field in a relational
30 database.

A Business Intelligence Charts Repository Component 41d serves two purposes in the database server system 41. A first purpose is to improve the performance of chart retrieval process. The chart repository 41d captures and stores the most frequently visited charts in a central location. When an end user requests a chart, system 41 first queries the chart repository 41d to see if there is an existing chart. If there is a preexisting chart, server 41e pulls that chart directly from the repository. If there is no preexisting chart, server 41e runs the corresponding predefined query from the query repository 41c in order to extract data from data mart 41b and subsequently feed the data to the requested chart. A second purpose is to allow chart sharing, collaboration and distribution among the end users. Because charts are treated as objects in the chart repository, users can bookmark a chart just like bookmarking a regular URL in a web browser. They can also send and receive charts as an email attachment. In addition, users may logon to system 41 to collaboratively make decisions from different physical locations. These users can also place the comments on an existing chart for collaboration.

Another component of system 41 is the Web Server component 41e, which has a number of subcomponents. A web server subcomponent (such as Microsoft IIS or Apache server or any other commercially available web servers) serves HTTP requests. A database server subcomponent (such as Tango, Cold Fusion or PHP) provides database drill-down functionality. An application server subcomponent routes different information requests to different other servers. For example, sales revenue

chart requests will be routed to the database system 41; customer profile requests will be routed to a Data Mining server, and competition information requests will be routed to a Text Mining server. The latter two systems
5 are discussed below. Another subcomponent of server 41e is the chart server, which receives requests from the application server. It either runs queries against data mart 41b, using query repository 41c, or retrieves charts from chart repository 41c.

10 As output 43, database server system 41 delivers business intelligence about an organization's sales performance as charts over the Internet or corporate Intranet. Users can pick and choose charts by regions, by quarters, by products, by companies and even by
15 different chart styles. Users can drill-down on these charts to reveal the underlying data sources, get detailed information charts or detailed raw data. All charts are drill-down enabled allowing users to navigate and explore information either vertically or
20 horizontally. Pie charts, bar charts, map views and data views are delivered via the Internet or Intranet.

As an example of operation of system 41, gross revenue analysis of worldwide sales may be contained in predefined queries that are stored in the query
25 repository 41c. Gross revenue queries accept region and/or time period as parameters and extract data from the Data Mart 41b and send them to the web server 41e. Web server 41e transforms the raw data into charts and publishes them on the web.

30

Data Mining System

FIGURES 6 and 7 illustrate a data mining system 61, which may be used within system 10 of FIGURES 1 and 2. FIGURE 6 illustrates the elements of system 61 and FIGURE 7 is a data flow diagram. Specifically, system 61 could be used to implement the profiles process 25, which collects profile data about the client.

Data sources 62 for system 61 are the Data Mart 41b, e.g., data from the tables that reside in Data Mart 41b, as well as data collected from marketing campaigns or sales promotions.

For data coming from the Data Mart 41b, data acquisition process 61a between Mining Base 61b and Data Mart 41b extract/transfer and format/transform data from tables in the Data Mart 41b into Data Mining base 61b. For data collected from sales and marketing events, data acquisition process 61a may be used to extract and transform this kind of data and store it in the Data Mining base 61b.

Data Mining base 61b is the central data store for the data for data mining system 61. The data it stores is specifically prepared and formatted for data mining purposes. The Data Mining base 61b is a separate data repository from the Data Mart 41b, even though some of the data it stores is extracted from Data Mart's tables. The Data Mining base 61b can reside in DB2, Oracle, Sybase, MS SQL server, P.SQL or similar database application.

Chart repository 61d contains data mining outputs. The most frequently used decision tree charts are stored in the chart repository 61d for rapid retrieval.

Customer purchasing behavior analysis is accomplished by using predefined Data Mining models that are stored in a model repository 61e. Unlike the predefined queries of system 41, these predefined models
5 are industry-specific and business-specific models that address a particular business problem. Third party data mining tools such as IBM Intelligent Miner and Clementine, and various integrated development environments (IDEs) may be used to explore and develop
10 these data mining models until the results are satisfactory. Then the models are exported from the IDE into standalone modules (in C or C++) and integrated into model repository 61e by using data mining APIs.

Data mining server 61c supplies data for the models,
15 using data from database 61c. FIGURE 7 illustrates the data paths and functions associated with server 61c. Various tools and applications that may be used to implement server 61c include VDI, EspressoChart, and a data mining GUI.

20 The outputs of server 61e may include various options, such as decision trees, Rule Sets, and charts. By default, all the outputs have drill-down capability to allow users to interactively navigate and explore information in either a vertical or horizontal direction.
25 Views may also be varied, such as by influencing factor. For example, in bar charts, bars may represent factors that influence customer purchasing (decision-making) or purchasing behavior. The height of the bars may represent the impact on the actual customer purchase
30 amount, so that the higher the bar is the more important the influencing factor is on customers' purchasing

behavior. Decision trees offer a unique way to deliver business intelligence on customers' purchasing behavior. A decision tree consists of tree nodes, paths and node notations. Each individual node in a decision tree
5 represents an influencing. A path is the route from root node (upper most level) to any other node in the tree. Each path represents a unique purchasing behavior that leads to a particular group of customers with an average purchase amount. This provides a quick and easy way for
10 on-line users to identify where the valued customers are and what the most important factors are when customer are making purchase decisions. This also facilitates tailored marketing campaigns and delivery of sales presentations that focus on the product features or
15 functions that matter most to a particular customer group. Rules Sets are plain-English descriptions of the decision tree. A single rule in the RuleSet is associated with a particular path in the decision tree. Rules that lead to the same destination node are grouped
20 into a RuleSet. RuleSet views allow users to look at the same information presented in a decision tree from a different angle. When users drill down deep enough on any chart, they will reach the last drill-down level that is data view. A data view is a table view of the
25 underlying data that supports the data mining results. Data Views are dynamically linked with Data Mining base 61b and Data Mart 41b through web server 61f.

Web server 61f, which may be the same as database server 41e, provides Internet access to the output of
30 mining server 61c. Existing outputs may be directly accessed from storage in charts repository 61d. Or

requests may be directed to models repository 61e.
Consistent with the application service architecture of
lead generation system 10, access by the client to web
server 61f is via the Internet and the client's web
5 browser.

Text Mining System

FIGURES 8 and 9 illustrate a text mining system 81,
which may be used within system 10 of FIGURES 1 and 2.
FIGURE 8 illustrates the elements of system 81 and FIGURE
10 9 is a data flow diagram. As indicated in FIGURE 8, the
source data 82 for system 81 may be either external and
internal data sources. Thus, system 81 may be used to
implement both the prospects system and profiles system
of FIGURE 2.

15 The source data 82 for text mining system 81 falls
into two main categories, which can be mined to provide
business intelligence. Internal documents contain
business information about sales, marketing, and human
resources. External sources consist primarily of the
20 public domain in the Internet. Newsgroups, discussion
forums, mailing lists and general web sites provide
information on technology trends, competitive
information, and customer concerns.

More specifically, the source data 82 for text
25 mining system 81 is from five major sources. Web Sites:
on-line discussion groups, forums and general web sites.
Internet News Group: Internet newsgroups for special
interests such as alt.ecommerce and
microsoft.software.interdev. For some of the active
30 newsgroups, hundreds of news articles may be harvested on
a weekly basis. Internet Mailing Lists: mailing lists

for special interests, such as e-commerce mailing list,
company product support mailing list or Internet
marketing mailing list. For some of the active mailing
lists, hundreds of news articles will be harvested on a
5 weekly basis. Corporate textual files: internal
documents such as emails, customer support notes sales
notes, and digital voice records.

For data acquisition 81a from web sites, user-
interactive web crawlers are used to collect textual
10 information. Users can specify the URLs, the depth and
the frequency of web crawling. The information gathered
by the web crawlers is stored in a central repository,
the text archive 81b. For data acquisition from
newsgroups, a news collector contacts the news server to
15 download and transform news articles in an html format
and deposit them in text archive 81b. Users can specify
the newsgroups names, the frequency of downloads and the
display format of the news articles to news collector.
For data acquisition from Internet mailing lists, a
20 mailing list collector automatically receives, sorts and
formats email messages from the subscribed mailing lists
and deposit them into text archive 81b. Users can
specify the mailing list names and address and the
display format of the mail messages. For data
25 acquisition from client text files, internal documents
are sorted, collected and stored in the Text Archive 81b.
The files stored in Text Archive 81b can be either
physical copies or dynamic pointers to the original
files.

30 The Text Archive 81b is the central data store for
all the textual information for mining. The textual

information it stores is specially formatted and indexed for text mining purpose. The Text Archive 81b supports a wide variety of file formats, such plain text, html, MS Word and Acrobat.

5 Text Mining Server 81c operates on the Text Archive 81b. Tools and applications used by server 81c may include ThemeScape and a Text Mining GUI 81c. A repository 81d stores text mining outputs. Web server 81e is the front end interface to the client system 13,
10 permitting the client to access database 81b, using an on-line search executed by server 81c or server 81e.

 The outputs of system 81 may include various options. Map views and simple query views may be delivered over the Internet or Intranet. By default, all
15 the outputs have drill-down capability to allow users to reach the original documents. HTML links will be retained to permit further lateral or horizontal navigation. Keywords will be highlighted or otherwise pointed to in order to facilitate rapid location of the
20 relevant areas of text when a document is located through a keyword search. For example, Map Views are the outputs produced by ThemeScape. Textual information is presented on a topological map on which similar "themes" are grouped together to form "mountains." On-line users can
25 search or drill down on the map to get the original files. Simple query views are similar to the interfaces of most of the Internet search engines offered (such as Yahoo, Excite and HotBot). It allows on-line users to query the Text Archive 81b for keywords or key phrases or
30 search on different groups of textual information collected over time.

A typical user session using text-mining system 81 might follow the following steps. It is assumed that the user is connected to server 81e via the Internet and a web browser, as illustrated in FIGURE 1. In the example
5 of this description, server 81e is in communication with server 81c, which is implemented using ThemeScape software.

- 10 1. Compile list of data sources (Newsgroups, Discussion Groups, etc)
2. Start ThemeScape Publisher or comparable application
- 15 3. Select "File"
4. Select "Map Manager" or comparable function
- 20 5. Verify that server and email blocks are correctly set. If not, insert proper information.
6. Enter password.
- 25 7. Press "Connect" button
8. Select "New"
9. Enter a name for the new map
- 30 10. If duplicating another maps settings, use drop down box to select the map name.
11. Select "Next"
- 35 12. Select "Add Source"
13. Enter a Source Description

14. Source Type remains "World Wide Web (WWW)"
15. Enter the URL to the site to be mined.
16. Add additional URLs, if desired.
17. Set "Harvest Depth." Parameters range from 1 level to 20 levels.
18. Set "Filters" if appropriate. These include Extensions, Inclusions, Exclusions, Document Length and Rations.
19. Set Advanced Settings, if appropriate. These include Parsing Settings, Harvest Paths, Domains, and Security and their sub-settings.
20. Repeat steps 14 through 20 for each additional URL to be mined.
21. Select "Advanced Settings" if desired. These include Summarization Settings, Stopwords, and Punctuation.
22. Select "Finish" once ready to harvest the sites.
23. The software downloads and mines (collectively known as harvesting) the documents and creates a topographical map.
24. Once the map has been created, it can be opened and searched.

Access to User Profiles Database

As explained above in connection with FIGURE 2, the profiles generation process 25 may be used to generate a profiles database 26. This database 26 stores

information about the client and its customers that may be used to better identify prospective customers.

Referring again to FIGURES 5, 7 and 9, various mining processes used to implement system 10 may access and use the data stored in database 26. For example, as illustrated in FIGURE 5, the database server 41e of database server system 41 may access database 24 to determine user preferences in formulating queries and presenting outputs. As illustrated in FIGURE 7, the data mining server 61c of data mining system 61 may access database 24 for similar purposes. Likewise, as illustrated in FIGURE 9, the text mining server 81c of system 81 may access database 24 to determine preferences in formulating queries, especially during query drill downs.

Text Indexing System

FIGURES 10 and 11 illustrate a text indexing system 101, which may be used within system 10 of FIGURES 1 and 2. FIGURE 10 illustrates the elements of system 101 and FIGURE 11 is a data flow diagram. Like system 81, system 101 may be used to implement either the prospects process 21 or profiles process 25 of FIGURE 2.

Text mining system 81 and text indexing system 101 are two different systems for organizing mass textual information. Text mining system 81 identifies and extracts key phrases, major topics, and major themes from a mass amount of documents. The text mining system 81 is suitable for those on-line users who want to perform thorough research on the document collection. Text indexing system 101 is similar to text mining system 81 but is simpler and faster. It only identifies and

extracts syntax information such as key words/key phrases. It provides a simple and fast alternative to users who just want to perform keyword searches.

The data sources 102 for Text Indexing system 101
5 are similar to those described above for Text Mining system 81. For data acquisition 101a, various software may be used. These include web crawlers and mailing list collecting agents. These are similar to those described above in connection with Text Mining system 81.

10 The text archive 101b is the central data store for all the textual information for indexing. The textual information it stores is specially formatted and indexed for text mining or indexing purpose. The Text archive 101b supports a wide variety of file formats, such plain
15 text, html, MS Word and Acrobat. Text archive 101b may be the same text archive as used in system 81.

Server 101c indexes the document collection in a multi-dimensional fashion. It indexes documents not only on keywords/key phases but also on contact information
20 associated within the documents. In other words, the server 101c allows on-line users to perform cross-reference search on both keywords and contact information. As an example, when users perform a keyword search on a collection of documents, the text indexing
25 server returns a list of hits that consist of relevance (who-when-what), hyperlink, summary, timestamp, and contact information. Alternately, when users perform contact information search on a collection of documents, the text indexing server 101c yields a list of documents
30 associated with that individual.

Using Text Indexing Server 101c, users may navigate documents easily and quickly and find information such as "who is interested in what and when."

Contact information and links to the associated
5 documents are migrated into a Sales Prospects repository 101d (a relational database). This contact information can be exported into normal contact management software from the repository 101d.

The outputs 103 of system 101 are varied. Simple
10 Query Views may be delivered to the client over the Internet or Intranet. By default, all the outputs have drill-down capability to allow users to reach the original documents. The Query Views may be similar to the interfaces of commonly used Internet search engines
15 offered, such as Yahoo, Excite and HotBot. It allows on-line users to query the Text Archive 101b for keywords/key phrases and contact information search on different groups of textual information collected over time.

20 FIGURE 11 illustrates the operation of text indexing server 115, which may be used to integrate queries from both text database 101b and another database 111 that stores information about prospective customers. For example, database 111 might be any one of the databases
25 26, 41b, 61b, or 81b of FIGURES 2, 4, 6, or 8. Server 115 accepts query parameters from the client, which may specify both contact parameters and keywords for searching database 111 and database 101b, respectively. The search results are then targeted toward a particular
30 category of prospects. FIGURE 11 also illustrates how server 115 may be used to store, identify, and reuse

queries. The queries for a particular client may be stored in user profiles database 26.

Digital Voice Recording Mining System

FIGURE 12 illustrates a digital voice recording mining system 120. System 12 may be used to implement the prospects process 21 of FIGURE 2, or it may be integrated into the text mining system of FIGURES 8 and 9.

Digital Voice Records (DVR) are increasing in use as companies move to sell and market over increasing boundaries, improve customer relations and provide a variety of support functions through call centers and third-party vendors. Present technology allows calls to be recalled through date-time stamps and a variety of other positional indicators but there are no means to analyze the content and context of the massive amount of this audio media.

System 120 uses speech-to-text translation capability to convert the digitally recorded voices, most often Vox or Wave (wav) format, into machine-readable text. A positional locator is created in the header file to facilitate direct linking back to the voice record, if needed. Accuracy of the recording on the receiving end is enhanced through training of the voice engine; an acceptable margin of error is expected on the incoming voice. The text files are stored in a Data Mart 122 where they may be mined using a search engine. Search engines such as ThemeScape are especially suitable in that they do more than simply count words and index frequently occurring phrases; they find "themes" by

examining where words appear in the subject, text and individual sentence structure.

A typical user session of system 120 might follow the following steps: Call is either received or
5 initiated. Depending on state law, the parties are advised that the call may be recorded for quality control purposes. Call is digitally recorded using existing technology from providers such as 1DigiVoice. Vox or Wave (voice) files 121 are translated using speech-to-
10 text conversion programs. Text files are stored in logical areas in Data Mart 122, for mining with a search engine. Maps or similar visual/graphical representations are placed in a Map or Image Repository 123. Users search maps using the search engines browser plug-in.
15 When the user finds documents to review, the user is prompted to select "voice" or "text." If text, the original document/file in the Data Mart is displayed in the browser window. If voice, the positional indicator is pumped to the Digital Voice Record application that
20 locates, calls and then plays to voice file segment.

Referring again to FIGURE 8, the voice data mart 122 may be one of the data sources for text mining system 81. Text mining server 81c is programmed to execute the functions of FIGURE 12 as well as the other functions
25 described above in connection with FIGURES 8 and 9. Similarly, the text in Data Mart 120 could be indexed using server 101c of FIGURES 10 and 11. In today's technological environment, the DVR storage 121 would originate from internal storage of the client, but
30 Internet retrieval is also a possibility.

Other Embodiments

Although the present invention has been described in detail, it should be understood that various changes, substitutions, and alterations can be made hereto without departing from the spirit and scope of the invention as defined by the appended claims.